Prediction of Sugarcane Yield Using Machine Learning Models

Introduction

This project aims to **predict sugarcane yield** using satellite-derived vegetation indices across a 12-month period. Two ensemble learning models—**Random Forest (RF)** and **Gradient Boosting Regressor (GBR)**—were implemented to train predictive models using a feature set of 72 remote sensing variables.

Dataset

The data used in this project was obtained from two primary sources:

- **MODIS (Moderate-Resolution Imaging Spectroradiometer)**: MODIS provides medium-resolution satellite data and offers approximately 44 geophysical products. The following specific MODIS products were used:
  - **MOD13**: NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index)
  - **MOD15**: LAI (Leaf Area Index), FPAR (Fraction of Photosynthetically Active Radiation)
  - **MOD16**: ET (Evapotranspiration)
  - **MOD17**: GPP (Gross Primary Productivity)
- **ICRISAT (International Crops Research Institute for the Semi-Arid Tropics)**: Provided the actual sugarcane yield data for 15 districts in Western Uttar Pradesh.

Vegetation and Productivity Indices

1. **NDVI (Normalized Difference Vegetation Index)**
   Indicates vegetation greenness and health using red and near-infrared reflectance. Values range from -1 to +1, where higher values indicate dense and healthy vegetation. It is widely used in crop monitoring and drought assessment.

2. **EVI (Enhanced Vegetation Index)**
   An improvement over NDVI, especially in areas with dense vegetation or high humidity. It uses blue reflectance to correct for atmospheric effects, providing more accurate vegetation estimates.

3. **LAI (Leaf Area Index)**
   Represents the total leaf surface area per unit ground area. It reflects canopy density and is essential for understanding photosynthesis, water use, and plant growth.

4. **GPP (Gross Primary Productivity)**
   Measures the total carbon fixed by plants through photosynthesis, indicating overall vegetation productivity and used in carbon cycle modeling.

5. **ET (Evapotranspiration)**
   Combines soil evaporation and plant transpiration to estimate total water loss, useful in understanding crop water requirements and irrigation management.

6. **FPAR (Fraction of Photosynthetically Active Radiation)**
   Measures the fraction of sunlight absorbed by plants for photosynthesis. It is directly related to GPP and reflects how efficiently plants convert sunlight into biomass.

Data Extraction and Processing

- The MODIS products were extracted using the **Application for Extracting and Exploring Analysis Ready Samples (AppEEARS)** developed by NASA (AppEEARS Team, 2020). AppEEARS allows efficient subsetting of geospatial datasets using spatial, temporal, and variable-specific filters.

- Requests were raised via AppEEARS for 15 districts in Western Uttar Pradesh: *Muzaffarnagar, Bulandshahr, Meerut, Saharanpur, Aligarh, Mathura, Agra, Mainpuri, Moradabad, Rampur, Bijnor, Bareilly, Etah, Shahjahanpur, Pilibhit, Badaun, Kasganj.*

- Yield data for sugarcane (in kg/ha) was collected district-wise from ICRISAT.

- Satellite-derived variables were collected monthly from April to March, resulting in 72 features (6 variables × 12 months). Each MODIS product was filtered to retain only the **mean** value per month per district to ensure consistency and reduce noise.

These indices are crucial for monitoring vegetation health, estimating crop yield, analyzing water use, and modeling ecosystem productivity.

Objectives

- Use 72 remote sensing features for predictive modeling.

- Train and evaluate Random Forest and Gradient Boosting Regressors.

- Compare model performances using R^2, MSE, MAE, and cross-validation.

Methodology

1. Data Preparation

- Feature matrix: 72 satellite-based variables.
- Target: Sugarcane yield.
- Train-test split: 80% training and 20% testing.

2. Model Training and Tuning

- **Random Forest Regressor:**
  - Best Parameters: `n_estimators=100, min_samples_split=2, min_samples_leaf=2, max_depth=10, bootstrap=True`
- **Gradient Boosting Regressor:**
  - Best Parameters: `n_estimators=100, min_samples_split=2, min_samples_leaf=1, max_depth=4, learning_rate=0.05`

3. Model Evaluation

- Performance metrics:
  - R^2 (coefficient of determination)
  - Mean Squared Error (MSE)
  - Mean Absolute Error (MAE)
  - Cross-validated R^2

4. Feature Importance

- Identified key features contributing to sugarcane yield prediction.
- FPAR and LAI ranked highest in both models.

Results

Model	$\mathbf{R^2}$	**RMSE**	**MAE**	**Cross-Validated** $\mathbf{R^2}$
Random Forest	∼0.52	∼588.69	∼467.19	∼0.65
Gradient Boosting	∼0.59	∼542.64	∼421.56	∼0.68

Note: Gradient Boosting slightly outperforms Random Forest in terms of all evaluation metrics.

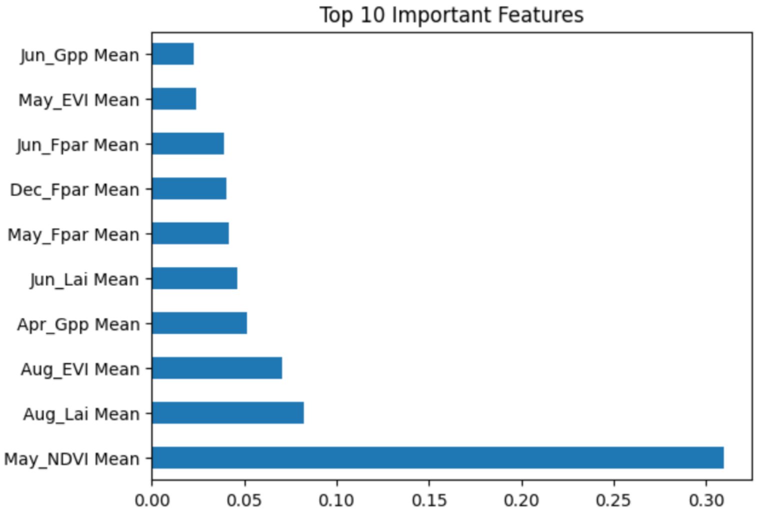

Figure 1: Top 10 Important Features — GBR Model

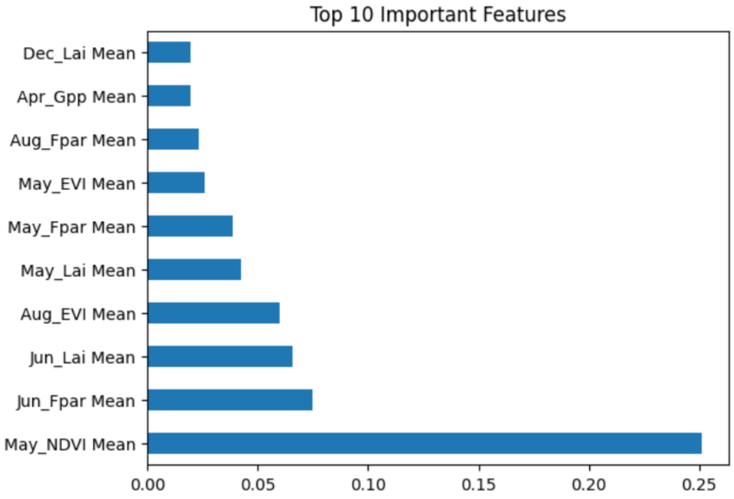

Figure 2: Top 10 Important Features — Random Forest Model

Conclusion

The sugarcane yield for 15 districts in Western Uttar Pradesh was estimated using satellite-based monthly vegetation indicators. Among the tested models, the Gradient Boosting Regressor achieved the best performance with an R^2 of 0.59 and RMSE of 542.64. The analysis confirms that remote sensing features, particularly FPAR and LAI, play a vital role in yield estimation as they directly relate to the photosynthetic capacity and biomass accumulation of the crop.